PREDICTIVE VALIDITY OF MATHEMATICS-CURRICULUM BASED MEASUREMENT

A Thesis
by
SARA BROWNING REYNOLDS

Submitted to the Graduate School
Appalachian State University
in partial fulfillment of the requirements for the degree
MASTER OF ARTS

May 2011
Department of Psychology

PREDICTIVE VALIDITY OF MATHEMATICS-CURRICULUM BASED MEASUREMENT

A Thesis
by
SARA BROWNING REYNOLDS
May 2011

APPROVED BY:

_____
Jamie Yarbrough Fearrington
Chairperson, Thesis Committee

_____
Sandra Gagnon
Member, Thesis Committee

_____
Pamela Kidder-Ashley
Member, Thesis Committee

_____
Rose Mary Webb
Member, Thesis Committee

_____
James C. Denniston
Chair, Department of Psychology

_____
Edelma D. Huntley
Dean, Research and Graduate Studies

FOREWORD

This thesis is written in accordance with the style of the *Publication Manual of the American Psychological Association (6th Edition)* as required by the Department of Psychology at Appalachian State University.

Predictive Validity of Mathematics-Curriculum Based Measurement

Sara Browning Reynolds

Appalachian State University

Abstract

As accountability, early screening, and prevention become more prevalent within schools, it is imperative that educators have knowledge of effective tools that screen for academic failure. Although mathematical competency is associated with life skills and economic success, research in student mathematical performance offers less substantial evidence when compared to investigations of student performance in reading. Empirical evidence to support the predictive validity of Mathematics-Curriculum Based Measurement (M-CBM) to identify accurately learners at-risk for low academic performance is a critical research area. This study analyzed correlations between M-CBM benchmark scores and student performance on the Tennessee Comprehensive Assessment Program (TCAP), a state mandated high-stakes test (i.e., related to federal and local funding, student placement decisions, teacher tenure, etc.). Participants were 1,732 students enrolled in grades 3-8 in a rural southeastern school system. Linear regression models were used to investigate the research questions. Specifically, this study sought to determine to what degree the Fall, Winter, and Spring M-CBM scores are correlated with the TCAP results. Additionally, the authors sought to assess the validity of M-CBMs from three time points for predicting TCAPs at the various grade levels. Lastly, the authors sought to identify any temporal differences in M-CBM and TCAP correlations based upon the time of the benchmarking (i.e., Fall, Winter, and Spring). Theoretical and instructional implications of the results as well as directions for future research in this area are discussed.

Predictive Validity of Mathematics-Curriculum Based Measurement

Mathematical proficiency is increasingly regarded as crucial to our nation's economy and essential for individuals to successfully complete tasks encountered in everyday life (Reyna & Brainerd, 2007). On April 18, 2006, as part of his agenda to "strengthen math education in order to give our students the skills to succeed in the 21st century," former President George W. Bush issued an Executive Order, creating the National Mathematics Advisory Panel (National Mathematics Panel, 2007). Research continues to support the need for a focus on improving the mathematical competencies of elementary and secondary students within the United States. The National Assessment of Educational Progress assessed mathematics achievement in a nationally representative sample of 168,000 fourth-grade students and 161,000 eighth-grade students. Their findings indicated that only 39% of fourth graders and 34% of eighth graders were at or above proficiency in mathematics in 2009 (National Center for Education Statistics, 2009). Improvements in mathematical proficiency should be a key issue for policymakers, given their link to positive life outcomes into adulthood (Reyna & Brainerd).

However, ensuring a free and appropriate education to all students does come at a high price. Sattler (2008) reported that the total cost of special education services in the United States during 1999-2000 was nearly $50 billion. Such figures point to the importance of research that informs educators about which programs and tools can most effectively and efficiently assist in improving mathematics outcomes for all students within this country. The purpose of the current study is to assess the accuracy with which

Mathematics-Curriculum Based Measurement (M-CBM) can predict student performance on a standardized measure of student mastery of grade-level mathematics curriculum.

Federal Education Initiatives, High-Stakes Testing, and Accountability

Given the cost of public education, it seems logical that government officials and taxpayers demand clear proof that educational programs are effective and worth the time and money required. Braden and Shroeder (2004) describe The No Child Left Behind Act (NCLB, 2001) as the most current reauthorization of the Elementary and Secondary Education Act (ESEA; P.L. 89-10), which was first enacted in 1965 as a part of the War on Poverty. The main focus of NCLB/ESEA is Title I, which provides funding to assist schools in educating economically disadvantaged children. Schools in which 40% or more of the student body is below the poverty line are eligible to receive these funds (Braden & Shroeder). Title I programs differ from special education programs in that schools eligible for Title I funds may serve any students with those funds. Title I services are highly important to NCLB legislation, because Title I requires state accountability for higher levels of student learning as measured through statewide testing (Braden & Shroeder). Adequate Yearly Progress (AYP) requires schools to demonstrate progress toward the goal of having 100% of students meet state proficiency standards on high-stakes tests by 2014. Dworkin (2005) details the consequences for schools that consistently fail to meet AYP, including redirection of a portion of the Local Education Agency's Title I funds to retrain teachers, consultation with outside educational experts, the option for students to leave the school, the removal of staff, and possible restructuring as a charter school.

As a result of the educational initiatives previously described, most school districts in the United States are currently conducting "high-stakes" assessment by gathering student performance data, mostly through local and state assessments (Sibley, Biwer, & Hesch, 2001). NCLB legislation leaves most of the design details of state accountability systems up to the states, and as a result many states have built upon the general measures of accountability that were adopted at earlier periods (Chubb, 2005). Furthermore, NCLB allows flexibility in the ways states measure AYP. Some states aggregate the results of high-stakes tests across two or three years, while others calculate AYP from a single year measure (Dworkin, 2005). These inconsistencies among states underscore the need for continued research investigating the effectiveness of accountability programs implemented in school districts around the country. Kelley (2008) claims "inconsistencies in math state standards, curricular focus, instructional delivery, and assessment practices are reasons for large numbers of students not demonstrating expected performance outcomes" (p. 419).

Opfer, Henry, and Mashburn (2008) studied districts' responses to high stakes accountability (HSA) in six southern states. They created policy profiles for Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee and used these profiles to describe state testing policies, professional development policies, and HSA policies. The investigators surveyed teachers about the systems implemented in their respective states during the 1999-2000 school year. The sample included 24 schools per state for a total of 144 schools. A 0-5 scale created by Carnoy and Loeb (2002) was utilized to assess the strength of accountability requirements applied by these six states. For example, sample states received an index score of 0 if they did not have statewide

testing or did not set statewide standards for schools and districts, whereas sample states received an index score of 5 if they had a high school exit exam and testing in elementary and middle grades with strong rewards or sanctions stipulated. Results indicated that few states made provisions for professional development in accountability or testing policies. Although five of the six states provided some financial support for professional development, North Carolina was the only state to finance most of the professional development provided to districts within the state. Regarding accountability policies, the results suggested variablity among the states. Whereas North Carolina and Kentucky were found to have been holding schools accountable for a decade, other states, such as Georgia and Mississippi, had just recently approved accountability policies. This study noted substantial differences in the amount and types of accountability structures in place.

The researchers also concluded that the level of HSA within a state was not associated with school support for teaching and learning, or for using assessment data. Systems with increased HSA indexes did appear to encourage district leaders to concentrate on teaching and learning to a greater extent than they would have without these systems (i.e., helping schools use information about student achievement to improve instruction, helping schools set benchmarks and evaluate progress toward school and district standards, promoting teacher leadership, helping schools develop and maintain high standards, etc.). Within this study, labeling (of school proficiency) was positively and significantly related to district support for teaching and instruction. Sanctions and rewards were negatively and insignificantly related to teacher's perceptions of more district involvement in teaching and learning. This study pointed out

the variability in how states and districts are implementing accountability policies and supports the need for continued research to examine the ways in which state responses to HSA policies may influence student learning and outcomes.

Comorbidity of Reading and Mathematics Deficits

NCLB mandated that the development of both reading and mathematics skills be a primary focus for schools. Such a joint focus seems justified given the comorbidity rates reported for deficits within these two domains. For example, using a large sample of students (N = 46,373) in the Chicago Public Schools, Grimm (2008) completed a longitudinal study that showed a positive relationship between third-grade students' reading comprehension achievement scores and the rate of change for three components of mathematical achievement in eighth grade. Three areas were assessed using tests from the Iowa Test of Basic Skills (Hoover, Dunbar, & Frisbie, 2005). These tests included Problem Solving and Data Interpretation (solving word problems and using tables and figures to obtain information, compare quantities, and determine trends); Math Concepts and Estimation (number properties and operations, algebra, geometry, measurement, probability, statistics, number sense, and mental arithmetic abilities); and Mathematical Computation (the use of addition, subtraction, multiplication, or division with whole numbers, fractions, decimals, and combinations of these types of numbers). Results were analyzed using a series of linear growth models and suggested that the three mathematical components changed linearly from third through eighth grade with substantial between-school and between-student differences in the intercept. These findings suggest that, from grade 3 through grade 8, the mathematics skills of students with greater reading capacity in third grade tended to change more rapidly, than did the

math skills of third-grade students with lower reading achievement. For example, students who had a higher level of reading comprehension in third grade tended to change faster in their problem solving and data interpretation skills than students with lower reading achievement in third grade (effect size of $\beta = .16$).

Although reading and mathematics are both emphasized in state curricula, the two domains receive substantially different resource allocations and attention. For instance, Grimm (2008) reported that the federal *Reading First* initiative cost $6 billion, whereas an initiative to improve mathematics instruction, *Science Excellence*, received $1 billion in funding. Other researchers also have noted the limited and narrow focus of the research examining mathematics assessment and intervention (Daly & McCurdy, 2002; Fuchs, Fuchs, & Hollenbeck, 2007; Gersten, Jordan, & Flojo, 2005). This evidence supports the need for continued research on state accountability programs, especially in mathematics.

IDEIA and Response to Intervention

When discussing accountability, it is important to point out, as did Braden and Tayrose (2008), that during the reauthorization of the Individuals with Disability Education Improvement Act (IDEIA, 2004) policymakers worked to align special education legislation with the mandatory accountability standards of NCLB. This alignment required states to ensure that students with disabilities are provided access to instruction in a general education setting, are expected to achieve the same proficiency in academics as non-disabled students, and are included in educational accountability efforts (Braden & Tayrose). Given the increased accountability required by NCLB and IDEIA, Cusumano (2007) reported that students who are not on track to meet identified goals

"must be identified early; at a point before the gap between expected outcomes and observed skills broadens…. [and] data must be used to identify why their learning trajectories are not progressing in the desired directions" (p. 24).

IDEIA allows educators to use either the traditional IQ-achievement discrepancy model or a Response to Intervention (RtI) approach to identify students at risk for a Specific Learning Disability. Educators and researchers engage in ongoing debate over which approach is best; Restori, Gresham, and Cook (2008) offered multiple reasons why many in the field advocate for the use of RtI. First and foremost, RtI relies on early screening and identification, which results in better intervention outcomes. Also, RtI utilizes assessment procedures that are directly linked to intervention, such as curriculum-based measurement (CBM), which is a research-based screening and progress monitoring tool. Furthermore, RtI demands the use of evidence-based interventions. Lastly, RtI moves educators from relying on the "wait to fail" ideology of the traditional discrepancy model and ultimately results in schools no longer relying on within-child explanations of learning disabilities.

Universal Screening, Progress Monitoring, and Curriculum-Based Measurement

A major goal of the RtI approach is to utilize screening tools that can proactively and accurately identify students in need of increased academic support, so that evidence-based interventions can be implemented and monitored for effectiveness. In order to implement a more strategic approach to monitoring student progress and responses to intervention, educators are using a three-tier problem solving model. Within this model, students are divided into three separate tiers based upon their scores on screening measures (Shinn, 2008). Tier 1 should encompass approximately 80% of the student

population. Educators are encouraged to gather benchmark scores for these students three to four times per year in order to monitor individual progress toward an annual goal. To implement Tier 1 instruction, a district must choose a core curriculum. This curriculum is presented in a whole-group or small-group format within the regular education classroom and is adapted to address standards and student needs identified by benchmark assessments (Brown-Chidsey, Bronaugh, & McGraw, 2009). Tier 2 should include approximately 15% of the student population. Educators should offer monthly strategic monitoring for students receiving intervention at this level. Tier 2 instruction includes small group instruction (three to six students) for two to three days per week. At this level, it is critical for interventions to be targeted toward skill deficits and matched to students' areas of need (Brown-Chidsey et al.). Lastly, Tier 3 should serve approximately 5% of the student population and offer the most intensive services, including weekly progress monitoring. Tier 3 instruction is offered in small groups (two to three students) or on an individual basis for five days per week. Intervention includes intensive, targeted instruction with multiple opportunities for students to respond and practice (Brown-Chidsey et al.).

Educators have learned that summative achievement tests utilized in the past are not satisfactory for monitoring student response to intervention because they are time consuming, unable to capture incremental skill changes, expensive, do not allow repeated administrations, and are not developed from instructional curricula (Salvia, Ysseldyke, & Bolt, 2007). According to Shinn (2008), "as schools move away from traditional systems of determining placement and services to systems with a problem-solving or solution-

focused orientation, the use of measurement procedures that can be administered efficiently and linked directly to intervention are required" (p. 245).

Public schools are increasingly utilizing CBM, along with other evidenced-based tools, as part of a three-tier problem solving model designed to assess the general student population and provide early intervention to children whose educational needs are beyond the scope of what the general curriculum can provide. In order to give teachers simple tools to write Individual Educational Program goals and monitor progress, Deno and colleagues at the University of Minnesota's Institute for Research on Learning Disability developed CBM in the mid 1970s (Deno, 2003). CBMs are 1- to 5-minute standardized tests used by educators in a general education setting to assess the effects of instructional interventions in the basic skills of reading, mathematics, spelling, and written expression (Shinn, 2008). CBM is a valuable tool in the RtI process, since it can be used to screen students for academic deficits; to create school, district, and national norms; to measure student achievement; and to reliably monitor progress toward goals (Jewell & Malecki, 2005). Merrell, Ervin, and Gimpel (2006) observed of CBM:

> These tools have demonstrated efficacy for direct assessment and monitoring of student academic performance within the curriculum. They provide an alternative to traditional norm-referenced assessment practices and have the advantage of being more closely tied to the curriculum, they are of shorter duration, they are sensitive to incremental changes, and they can be used repeatedly to monitor growth formatively. (p. 147)

In contrast, high-stakes tests are summative, meaning they only yield end-of-year scores that teachers often do not even see until the summer or at the beginning of the next

school year. Since CBM is formative, these measures give educators the ability to assess student progress and needs throughout a school year. Additionally, researchers have concluded that CBM may provide a more objective and accurate basis for determining which students are at academic risk than teacher report (Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006).

Research Regarding the Effectiveness of CBM

To assist in achieving the ultimate goal of increasing the number of students who score at the proficient level on high-stakes tests, best practice recommends incorporating CBM within RtI models; CBM can be used proactively to screen for skill deficits and thereby facilitate early intervention to students in need. Given this emphasis, researchers should continue to investigate the effectiveness of CBM for academic screening. As indicated by Wallace, Espin, McMaster, Deno, and Foegen (2007), "The breadth and depth of CBM research varies….Substantial research has been conducted in the elementary grades; less has been conducted in the secondary grades. Reading has received more attention than has mathematics" (p. 66). Despite the lack of attention given to CBM, and mathematics in particular, the literature offers a plethora of research supporting the predictive validity of CBM, especially in the subject of reading.

Hintze and Silberglitt (2005) followed 1,766 students from first through third grade, utilizing Reading-Curriculum based measurement (R-CBM) benchmark assessments completed three times within a school year. Students also were administered the Minnesota Comprehensive Assessment (Minnesota Department of Education, 2003) at the end of third grade. These researchers studied predictive validity by analyzing the R-CBM cut scores (scale scores that separate and define performance levels) comparing

three statistical procedures: discriminative analysis, logistic regression, and receiver operator characteristics curves. Minnesota Comprehensive Assessment (MCA) reading cut scores of 1420 and above were considered passing, and scores below 1420 were failing. R-CBM cut scores for each benchmarking period were determined using student reading MCA performance as the criterion standard. Then a cut score for R-CBM for Spring of third grade was initially determined using reading MCA performance as the criterion standard. Once this was set, the third grade Spring cut score was the criterion standard for determining the third grade Winter cut score. This process continued in this sequentially backward process with each R-CBM benchmark used to determine the cut score for the benchmark that occurred immediately before. The four possible outcome proportions that were assessed from a diagnostic accuracy analysis were sensitivity, specificity, positive predictive power, and negative predictive power. Sensitivity refers to the probability that when a diagnostic status is present in the criterion, the individual will be identified positively by the predictor. Specificity describes the probability that when a diagnostic status is absent on the criterion, the individual will not be identified by the predictor. Positive and negative predictive powers are measures of efficacy and reflect the probability that a predictor measure will correctly discriminate between who will or will not be identified by the criterion measure, respectively. Correlations reported ranged from .49 to .94. The investigators stated that "each statistical procedure investigated set cut scores that yielded adequate levels of both diagnostic accuracy and efficiency" (p. 382).

Based on their findings, Hintze and Silberglitt suggested that R-CBM was highly correlated with MCA performance at all grade levels and was accurate and efficient in

predicting which students were likely to pass the reading section of the MCA beginning in first grade. Not surprisingly, the results indicated that R-CBM was more strongly correlated with the MCA when the two assessments were collected closer in time. This study extends previous findings supporting the use of R-CBM as a significant predictor of broader measures of reading abilities.

Similarly, McGlinchey and Hixson (2004) investigated the predictive value of R-CBM for performance on the Michigan Educational Assessment Program's (MEAP) fourth grade reading assessment. This study spanned eight years and included 1,362 fourth-grade general and special education students. The R-CBM probes utilized were three passages randomly selected from the district basal fourth grade reading text, the Macmillan Connections Reading Program (Arnold & Smith, 1987). All students were administered the same reading passages in the 2 weeks before administration of the MEAP. One hundred words correct per minute (WCPM) was selected as the cut score for the R-CBM passages. Individual student data were analyzed and diagnostic efficiency statistics were used to determine the accuracy of the reading rate cut score. Five statistical measurements were used to determine diagnostic accuracy; they included Sensitivity (the percentage of students who failed the MEAP and who read less than 100 WCPM); Specificity (the percentage of students who passed the MEAP and who read 100 WCPM or greater); Positive Predictive Power (the probability that a student reading less than 100 WCPM would score less than satisfactory on the MEAP); Negative Predictive Power (the probability that a student reading greater than or equal to 100 WCPM would score satisfactory on the MEAP); and the Overall Correct Classification (the percent of agreement between WCPM cut scores and MEAP performance). Results

indicated that specificity for identifying students who achieved satisfactory scores on the MEAP was 74%, and the sensitivity for identifying those who did not achieve satisfactory scores was 75%. The positive predictive power of the cut score was 77%, and the negative predictive power was 72%. The overall correct classification was reported to be 74%. Cohen's kappa was .48, meaning that the diagnostic efficiency of the R-CBM cutoff was 48% above chance. This study indicates a moderately strong relationship between oral reading rates and performance on the MEAP and adds to the literature supporting R-CBM probes as a valid and reliable assessment of reading skills.

In a study conducted by Crawford, Tindal, and Stieber (2001), a CBM of reading aloud from narrative passages was used to predict performance on statewide achievement tests in the areas of reading and math. To assess students' reading rates, three passages from the Houghton Mifflin Basal Reading Series (1989) were modified to contain approximately 200 to 250 words. Both math and reading performance were addressed in this study because the math multiple-choice achievement tests required proficient reading skills. The researchers provided longitudinal data for students across a two-year period that included second to third grade (n = 77 in the first year and n = 51 in the second year.) During the second year, students were tested on statewide criterion-referenced tests containing multiple-choice questions and performance tasks measuring math and reading proficiency (Oregon Department of Education, 1999). Three types of outcomes were reported. First, descriptive statistics were reported for Year 1 and Year 2. Second, correlations between the timed oral reading and the statewide reading and math tests were reported. Third, a chi-square analysis was used to determine which levels of oral reading rates were most predictive of performance on the statewide tests.

Results showed that the mean for scores on the statewide reading assessment met the state-established cut score for passing (scale score = 201). However, the mean scores on the statewide math assessment were below the criterion by 2 points (scale score = 202). Out of the 51 students with scores reported, representing all students in the study, 65% passed the reading assessment and 45% passed the math assessment. The mean gain in oral reading rate was approximately 42 correct words per minute. A Pearson correlation coefficient was calculated between second and third grade oral reading rates and indicated a strong relationship ($r = .84$, $p = .001$).

Lastly, chi-square analyses were utilized and presented as a 2 × 4 classification table for the within-year scores and another for the across-years scores. The within-year data were reported using norms established by Hasbrouck and Tindal (1992). Results indicated that students reading below the 25th percentile in the Winter of third grade read between 0 to 70 correct words per minute (CWPM). These rates were used to establish the first cell. The remaining three cells modeled quartiles in Hasbrouck and Tindal's study and used the following rates: 71 to 92 correct CWPM for the second cell; 93 to 122 CWPM for the third cell; and 123 or more CWPM for the fourth cell. The strongest finding was that 81% of students reading at the third and fourth quartiles passed the statewide assessment, whereas only 37% of students reading at the first or second quartiles passed ($X^2 = 12.8$, $p = .005$). The across-years data revealed that students reading below the 25th percentile in the Winter of second grade read between 0 to 46 words per minute. The second cell was represented by 47 to 77 CWPM, the third cell as 78 to 105 CWPM, and the fourth cell as 106 or more CWPM. Of the 37 students reading in these top 3 quartiles, 29 passed the statewide reading test (78%), whereas only 29% of

students reading in the first quartile passed ($X^2 = 16.8$, $p = .001$). By finding a strong association between timed oral reading rates for students in second grade and their reading rates in third grade, this research helped to support the stability of CBM as a tool to assess performance across different populations of students.

VanDerHeyden, Witt, Naquin, and Noell (2001) added to the research investigating the role CBM can play in early intervention. These researchers developed a series of six group-administered CBM probes to assist in the identification of kindergarten students showing deficiencies in school readiness skills. The *Circle Number Probe* required students to count a set of circles on one side of a page and to circle the correct number from a list of possible choices on the other side of the page. The *Write Number Probe* required students to count a set of objects and write that number in a corresponding box. In the *Draw Circles Probe* students were required to draw in the space on the right hand side of the page the number of circles corresponding to the number specified in the left hand side of the page. The *Circle Letter Probe* presented students with a series of pictures, and each picture was followed by a row of four letters. Experimenters stated the name of each picture in five-second intervals, and students were instructed to circle the letter corresponding to the beginning letter sound of the picture name. The *Copy Letter Probe* presented students with capital letters ranging from A to Z that were arranged in ascending and descending order. Each letter was positioned over an empty box, and students were instructed to copy each letter in the box. Lastly, the *Discrimination Probe* presented four items (e.g., letters, numbers, shapes), three that matched and one that did not, arranged in a row. Students were instructed to circle the item that was different from the other items. Participants included 107 students from six

classrooms in two suburban schools located in south Louisiana. The findings for this study indicated acceptable alternate-form reliability for three of the six probes (r = .81 to .84). Scores on the six CBM probes were compared to ten subtests on the Comprehensive Inventory of Basic Skills, Revised (CIBS-R; Brigance, 1999), as well as the *Onset Recognition Fluency* subtest from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 1996; Kaminski & Good, 1996). The *Circle Letter Probe* correlated highest with the *Onset Recognition Fluency Probe* from DIBELS (r = .72). The *Circle Letter Probe* also correlated significantly with the CIBS-R (r = .68). All math readiness probes correlated moderately with math composite scores on the CIBS-R (.61, .44, .56, respectively). The investigators also used a discriminant function to determine if scores on the probes predicted retention. Results suggested that scores on probe measures accurately predicted students who would be retained in 71.4% of cases, who would not be retained in 94.4% of cases, and accounted for 77% of variance. Kappa was computed at .9 (p < .000) indicating strong prediction. These findings reinforce the literature that supports the use of CBM probes as screening devices to be used within an RtI model to inform early intervention.

Mathematics-Curriculum Based Measurement

Hosp, Hosp, and Howell (2007) outline the main characteristics of Mathematics-Curriculum Based Measurement. M-CBM probes are simple to administer and score and can be administered individually or with a group (Hosp et al.), qualities that make them ideal for use within RtI models. Target areas include Early Numeracy, Computation, and Concepts and Applications, although the specific scope can vary based on the curriculum

in use. Hosp et al. describe the measures within these three target areas, as presented below.

There are five separate Early Numeracy measures that may be utilized. Each of these probes must be administered individually and each requires one minute for administration. *Missing Numbers* requires students to tell the examiner the number that correctly completes a pattern represented by three other numbers. In the *Number Identification* measure, the student is presented with a sheet of numerals in random order and must state what each numeral is. For the *Oral Counting* measure, the student simply counts orally, starting at one. The *Quantity Array* measure presents the student with a box containing several dots, and the student must identify how many dots are in each box. Lastly, the *Quantity Discrimination* measure presents the student with two adjoining boxes, each containing a number, and prompts the student to identify which number is greater.

Computation probes contain either single- or multiple-digit problems using addition, subtraction, division, or multiplication facts. The mathematics problems administered on a particular M-CBM computation probe should represent the skills a student is expected to master throughout an entire school year for his or her particular grade. For example, a third grade curriculum might include multi-digit addition and subtraction with and without regrouping and multiplication facts that include factors up to nine. Probes are scored by counting the total number of correct digits written. The number of correct digits in the solution to the problem is utilized (rather than the number of correct problems), because this measure is more sensitive to change, which is essential for progress monitoring. Probes can be administered individually or to a group and

usually have a two-minute duration. Computational fluency has traditionally been the focus of most math research and is the area assessed in the current study.

M-CBM has expanded to include other math skills, such as the areas measured through Concepts and Applications measures. These M-CBM sheets include math skills such as measurement, graph interpretation, time, estimation, and others commonly found in mathematics curricula. These skills are more complex than simple computation skills, and administration requires from six to eight minutes. The response format for these measures varies, as some are fill-in-the-blank and others are multiple-choice. Also, the first grade measure is read to the student, but all others are completed independently.

Reliability and Validity of M-CBM

Although more research has investigated the psychometric properties of R-CBM, researchers are steadily contributing to the body of evidence supporting the use of M-CBM. Thurber, Shinn, and Smolkowski (2002) examined reliability and validity using a confirmatory factor analytic approach to determine what constructs M-CBM actually measures. Three models were tested: a unitary model where Computation and Applications comprise a general math competence construct that M-CBM measures accurately; a two-factor model where Computation and Applications are distinct constructs and M-CBM is a measure of Computation; and a two-factor model where Computation and Applications are distinct and M-CBM is a measure of Applications. The M-CBM measure consisted of a range of computation skills including basic addition, subtraction, multiplication, and division facts. The findings provided evidence of high alternate-form reliability for M-CBM with a median correlation of .91 among the three given forms. Convergent validity was found, as the M-CBM correlated highly with other

measures of basic facts computation (median *r* = .82) and more moderately with commercial measures of math computation, such as the Stanford Diagnostic Mathematics Test and the California Achievement Tests (median *r* = .61). Performance on M-CBM also was less correlated to tests measuring math application (median *r* = .42). Results of model testing indicated the most defensible model was a two-factor model of mathematics assessment where Computation and Applications were distinct, though highly related constructs (*r* = .83). This evidence supports the continued use of M-CBM in the public schools as a measure of math constructs presented. This study also indicated that reading may be an important component of overall math competence, as R-CBM Maze (reading comprehension) was reported to correlate highly with the M-CBM computation (with correlations ranging between .57 to .92). Strong correlations were also suggested between R-CBM Maze and math fact probes containing a combination of addition, subtraction, multiplications, and division facts, with reported correlations ranging from .59 to .92. Although there are some limitations to this study, including low interscorer agreement and a sample of primarily Caucasian participants, the authors succeeded in adding to the literature demonstrating M-CBM to be reliable and valid for use in some skill areas. The findings also raise a question regarding the extent to which reading ability may play a role in math skills.

Keller-Margulis, Shapiro, and Hintze (2008) studied the relationship between benchmark assessments (basic skills data gathered from administering CBM probes to students during the Fall, Winter, and Spring terms within a school year) and the amount of growth within a year for reading, math computation, and math concepts and applications CBM, and a statewide achievement test. Participants ranged from grades 1

through 5 and came from six elementary schools located in eastern Pennsylvania. The total sample included 1,461 students in the reading group and 1,477 in the math group. The researchers used AIMSweb (2002) probes to measure oral reading fluency; Monitoring Basic Skills Progress-Math Computations (Fuchs, Hamlett, & Fuchs, 1998) probes consisted of a single sheet of 25 problems of mixed operations. Monitoring Basic Skills Progess-Math Concepts and Applications (Fuchs, Hamlett, & Fuchs, 1999) probes also were administered for grades 2-5, including 18 problems designed to assess whether students had mastered Concepts and Application skills expected for their grade level. Specifically, the Math Concepts and Applications measures addressed counting, number concepts, names of numbers, measurement, charts and graphs, money, fractions, applied computation, and word problems. These scores were compared to scores on the Pennsylvania System of School Assessment (PSSA), the measure of accountability requirements in Pennsylvania, and also to The TerraNova Achievement Test-Second Edition, with the aim of providing evidence for the validity of CBM. Benchmark data was reportedly collected over a 10-15 day period during October, February, and May. The PSSA was administered in grades 3 and 5 in the Spring of the school year 1 and 2 years after normative comparison data were collected. The TerraNova was only administered to fourth grade students. Results indicated that the CBM data were moderately and positively correlated with the statewide achievement test and with the nationally normed instrument. This study supports the strength of CBM as a predictor of later performance on high-stakes tests.

Helwig, Anderson, and Tindal (2002) also investigated the ability of a CBM mathematics measure to predict scores on a statewide achievement test. In this study,

eight school districts within a western state were invited to participate in a pilot project to develop and test a series of CBM measures in reading, writing, and mathematics for grades 3, 5, 8, and 9. Only results for eighth-grade mathematics were reported for this study. All students (n = 171) were given a CBM math task along with the Computer Adaptive Test of Math Achievement (CAT), which served as an alternative for state achievement test results. Ninety students represented a general education population and 81 students were from special education. Pearson product-moment correlations were calculated between students' total number correct on the CBM math concept task and the corresponding CAT scores. Regression analysis was used to identify which combination of items most effectively predicted achievement. Results indicated that the performance of general education students was significantly higher than that of the special education students on both the CAT and CBM measures. The mean of correct answers on the CBM was 5.57 for the general education students and 1.77 for the special education students. A strong correlation emerged between the CBM and CAT for the general education students (r = .83) and a moderate correlation between the two for the special education students (r = .61). Using Discriminant Function Analysis, researchers were able to predict with 87% accuracy whether students would meet state mathematics standards. This study also is important in that it identifies CBM as an effective tool to monitor progress not only for general education but also for a special education population.

Foegen and Deno (2001) conducted a study to determine if M-CBM measures were potential indicators of growth in mathematics at the middle school level. Correlation and regression analyses were used to investigate the reliability and criterion validity of four measures of mathematics. One hundred students in the seventh and

eighth grades from an ethnically diverse middle school in an urban district acted as participants. Approximately 9.9% of these students were receiving special education services. The four measures studied were Basic Math Operations Task (BMOT), Basic Estimation Task (BET), and Modified Estimation Tasks (METs) presented in two forms, A and B. It should be noted that these CBM measures were created by the investigators for this study. The measures ranged from 1-3 minutes in duration. On two occasions during week 1 of the Spring of 1995, students completed a series of the assessment tasks. Scores for the measures were the number of correct responses made. The criterion variables included Math Grade Point Average (GPA), defined by the grade for the first semester of the school year, and subtest scores on the California Achievement Test (CAT). Mean scores on the BET reflected incremental increases across grades; this was not the case for the other measures. Lack of familiarity with the estimation measures was cited as a possible reason.

Internal consistency reliability coefficients ranged from .77 to .93; test-retest correlations ranged from .67 to .88; and parallel forms ranged from .67 to .86. Correlation coefficients for the scores obtained from the four measures and the criterion variables also were reported. Correlations with Math GPA were in the low to moderate range (.22 - .44). Moderate correlations were consistently observed with CAT subtest scores (.29 - .63). For the CAT scores, BMOT was the strongest predictor of performance on the Computation subtest, accounting for 44% of the variance. The BMOT and the BET both predicted scores on the Concepts subtest equally well but accounted for a smaller proportion (32%) of the variance. The BMOT best predicted Proficiency and Reasoning scores and accounted for slightly over 40% of the variance in

each rating. This study adds to the literature indicating that M-CBM measures are reliable and promising indicators of mathematics proficiency.

Fuchs et al. (2007) acknowledged that two types of errors challenge the accuracy of methods for classifying children into at-risk or not-at-risk groups: there are false positives, in which children who score below the cut-off on a predictive instrument and are labeled at risk later display academic competence on other tests; and false negatives, in which children score above the cut-off on a predictive instrument and later demonstrate academic difficulties on other tests. Using CBM screeners to assess Math Disability, Fuchs et al. assessed the predictive utility of these tools at the end of the second grade. They also examined the discriminant validity of math progress using four monitoring tools (M-CBMs). The four screeners incorporated a limited set of skills, such as number identification and counting, a more difficult single-skill screener that relied on fact retrieval, a multiple-skill computation screener that sampled the entire first-grade curriculum, and a multiple-skill concepts and applications screener that also sampled the entire first grade curriculum. The researchers administered these four screening measures to 170 students during September of first grade. Logistic regression was used to predict membership in the second grade Mathematics Disability and Non-Disability groups. For specifying the risk for Math Disability, the four-variable model demonstrated an AUC (area under the ROC curve) of .847 for Math Disability-Calculation and .806 for Math Disability-Word Problems. For predicting Math Disability-Calculation status, the four-variable screening model resulted in a hit rate of 78.2% and a rate of 74.7% for predicting Math Disability-Word Problem status. However, results also indicated that these four screeners would have resulted in 30 students being unnecessarily tutored (false positives)

and would have missed seven students (false negatives) who went on to meet Math Disability standards. This study points out that multi-skill screeners may correlate with various math outcome measures with varying degrees of strength. However, these findings do provide information to support the use of CBM Computation to measure math competence across first grade.

Hintze, Christ, and Keller (2002) focused on the generalizability of M-CBM single-skill and multiple-skills probes. The sample included 67 students enrolled in 21 first through fifth grade classrooms from an elementary school located in the Northeast. Students were administered grade-specific single-and multiple-digit calculation probes in a group testing arrangement, for a total of three probes per individual. Using a generalizability analysis (ANOVA), the authors investigated differences based on probe types and grade. Results from the tests of main effects revealed no significant differences between single-skill math probes (p = .61) or average performance across grades (p = .09), indicating single- and multiple-skill probes measure two distinct constructs. Results from a test of main effects found no significant differences between the three different single skills probes used during the CBM mathematics assessment (p = .61) or average performance across grades (p = .09). Multiple-skill math probes demonstrated more variability attributed to developmental or grade difference across students. Although this study leaves unanswered questions regarding the generalizability of multiple-skill M-CBM probes, the findings indicate that both single- and multiple-skill M-CBMs did measure distinct constructs and suggest a high level of dependability for single-measure probes in making educational decisions for children indentified as at-risk for academic difficulties.

Goals of this literature review include clarifying current information about mathematical difficulties seen within students and M-CBM's role in early identification of those students in need of support to improve their basic math skills. This topic is significant and warrants further investigation, given the poor mathematics performance of American students in comparison to those in other countries (National Mathematics Panel, 2007). In order to increase the mathematical proficiency of our students, researchers should continue to investigate the tools that best serve as early indicators of intervention needs in mathematics. Results of this review provide empirical support for the use of M-CBM as a screener for academic needs and evidence that M-CBM is a reliable and valid tool for measuring academic progress and is capable of predicting performance on high-stakes tests. A last goal of this review is to clarify and reiterate the need for more intensive research regarding mathematics, especially as compared to reading. Mathematics deserves research attention since mathematics proficiency is viewed as important to our nation's economy and represents skills crucial for individuals to complete daily life tasks (Reyna & Brainerd, 2007).

Purpose of the Current Study

The primary goal of this study was to investigate the predictive validity of M-CBM benchmark scores in reference to high stakes-test results. The author's intent was to provide research to support the use of M-CBM as a screening tool to assess the need for intensive academic support for children who are at-risk of scoring below proficiency on high-stakes tests. Two specific research questions were addressed in this study:

1. To what degree are the Fall, Winter, and Spring M-CBM scores correlated with Tennessee Comprehensive Assessment Program (TCAP) Mathematics

composite scaled scores at each grade level?

2. Are there temporal differences in M-CBM and TCAP correlations (i.e., are the Spring benchmark scores better predictors of TCAP performance, as compared to the Winter or Fall benchmark scores?)

Hypotheses

With regard to the first research question, it was predicted that, at a minimum, there would be moderate correlations between M-CBM and TCAP scores. This prediction was based on previous investigations, which found M-CBM scores to be adequate predictors of proficiency on other high-stakes tests (Helwig et al., 2002; Keller-Margulis et al., 2008). Next, we expected M-CBM scores from the earlier grades to be more strongly correlated with TCAP results because previous researchers have found M-CBM to be more highly correlated with high-stakes test results within the earlier primary grades (Foegen, 2008). Lastly, we predicted that the Spring benchmark scores would be more strongly correlated with TCAP scores than would the Fall and Winter benchmark scores, since both measures were administered during the Spring of the 2006-2007 school year. Prior evidence supports higher correlations between Spring benchmarks and high-stakes tests, as compared to Fall and Winter benchmarks (Keller-Margulis et al., 2008).

Method

Participants

The participants in this study were 1,732 students (51% boys and 49% girls) enrolled in general and special education classrooms in third through eighth grades. Participants attended five schools in a rural southeastern district, including two elementary schools, one intermediate school (grades 3-5), and two middle schools. See

Tables 1 and 2 for additional sample characteristics. The sample district enrolled 5,550 students and included a total of twelve schools, serving students in Kindergarten through grade 12. The school district granted permission to utilize these data for research purposes on May 13, 2008 (see Appendix A for approval). The current study gained university Institutional Review Board approval on June 2, 2009 (see Appendix B) and was conducted in accordance with ethical standards.

Materials and Procedure

AIMSweb (2002) M-CBM probes were administered to all students enrolled in grades 3 through 8. For each grade, the probes were comprised of computational problems representative of an annual grade level curriculum. A prototype grade-level M-CBM probe, constructed for each grade, arranged the order of the types of problems so that each probe would have an identical set of ordered problems (i.e., if the third grade M-CBM prototype had a basic addition fact such as (3+2) as the first problem, all third grade M-CBM probes would begin with a basic addition fact problem). A sample probe is included in Appendix C. The administration and scoring of M-CBM probes is a standardized process that includes clearly outlined procedures.

Shinn (2004) provides detailed procedures for administering and scoring M-CBM probes, which can be administered to students individually, in small groups, or class-wide, with the examiner carefully monitoring student participation. Examiners scored students' probes using the scoring method for Correct Digits (CD) that is specified for their grade and mathematics instruction approach. When scoring, examiners used answer keys that were provided, underlining the correct digits the student wrote and summing the total number of underlines. There are two scoring methods for determining correct digits.

For grades 1 through 4, counting the number of underlines in the answer only is the recommended method. Students were given credit for correct digits written, regardless of the full completion of the problem or if the problem was answered but crossed out or obviously reversed. For M-CBM probes in grades 5 through 8, examiners chose to score the number of underlines for answer-only or chose to score both the number of underlines and the critical processes used to obtain the answer. The answer-only method may be chosen when the curriculum teaches students various methods to solve computational problems. For this investigation, the number of CD for answer-only was utilized for all participating grades. Since these were benchmark measures given across multiple schools, resources (i.e., time, training) were of concern, and the investigators chose the shorter and more straightforward scoring approach. It was also considered that, although not all students show their work on screening measures during benchmarking, they can be instructed to do so if they are identified later for progress monitoring.

Use of the answer and critical processes method assumes there is a common way students have learned to solve more difficult computational problems. These more challenging problems are determined to be more valuable in terms of student outcomes, and they result in a higher CD score. When this method is used, the examiner uses an answer key that specifies which digits are to be counted. Each problem has an assigned CD value determined by what AIMSweb (2002) authors believe to be the most common way to solve the problem. When using this scoring method, students can receive CD scores not only for correct answers, but also for correctly writing the digits involved in the process of solving the computational problem.

Shinn (2004) offers a summary of studies to support the reliability of AIMSweb M-CBM probes. Thurber and Shinn (2002) indicated interscorer agreement of .83 and alternate form reliability of .91. Hintze et al. (2002) added to the literature supporting the construct validity of single and multiple skills CBM mathematics assessments, as evidenced by finding dependability coefficients of greater than .95. Researchers also have investigated convergent validity within this domain. M-CBM also has been found to correlate highly with other measures of basic facts computation (median $r = .82$) and to a lesser degree with commercial measures of math computation (median $r = .61$; Thurber et al., 2002).

The Tennessee Comprehensive Assessment Program (TCAP) Achievement Test assesses and reports student performance for K-12 education in the state of Tennessee. After an annual update, students take the TCAP each Spring to measure their basic skills in reading, language arts, mathematics, science, and social studies. This assessment directly measures skills that are included in the Tennessee state curriculum in all of these academic areas. The TCAP utilizes multiple choice questions and has set time limits. There are norm-referenced score interpretations for grades K-2 and criterion-referenced score interpretations for grades 3-8, as reported by the State of Tennessee Department of Education (2007). Students do not pass or fail; they are instead rated on a proficiency scale ranging from 1-5. A proficiency score of 1 indicates the student is lacking the basic academic skills expected for his or her grade level in that particular area. A proficiency score of 2 indicates the student is progressing. A proficiency score of 3 indicates the student is nearing proficiency. A proficiency score of 4 indicates the student is proficient in that particular academic area. Lastly, a proficiency score of 5 indicates the student is

advanced compared to other students in his or her grade level. Assessment results also include scaled score ranges. The scaled score ranges for each level are as follows: Below Proficient: 310-447 points; Proficient: 448-483 points; Advanced: 484-630 points. The TCAP was administered by classroom teachers during a one-week period in April 2007.

Procedure

The data for this study were collected as part of routine academic screenings. All M-CBM probes were administered and scored by members of a trained Student Assessment Team that included classroom teachers, teaching assistants, district school psychologists, and school psychology graduate students. The probes were administered during benchmarking in the months of August 2006, January 2007, and May 2007. Prior to the beginning of the study, training on AIMSweb M-CBM administration and scoring was completed by all members of the benchmarking team. To increase the reliability of results, raters were responsible for administering and scoring probes for a specific classroom across the three benchmark periods.

Per standardized training, the standardized directions for each probe were read aloud to the student(s), and examiners monitored and used prompts to ensure students did not skip around or excessively cross out problems they knew how to complete. Examiners did not provide corrections or feedback to students about the accuracy of their work during the testing. Probes were administered for a 2-minute (grades 1 through 3) or 4-minute (grades 4 through 8) duration, depending upon grade level, which was accurately tracked with a stopwatch or timer.

All members of the assessment team followed standardized administration instructions, reading aloud the following to the students:

> We're going to take a 2 (or 4) minute math test. I want you to write your answers to several kinds of math problems. Look at each problem carefully before you answer it. When I say begin, write your answer to the first problem and work across the page. Then go to the next row. Try to work each problem. If you come to one you really don't know how to do, put an 'X' through it and go to the next one. If you finish the first side, turn it over and continue working. Are there any questions? (pause) Begin.

Examiners also followed guidelines for prompting to address excessive skipping of problems by stating "Try to work each problem. You can do this kind of problem so don't skip or put an 'X' over it." If a student failed to work across the page, the examiner prompted him or her to "Work across the page. Try to work each problem in the row." Lastly, if a student stopped working before the test was completed, the examiner informed him or her to "Keep doing the best work you can." A full example of these standardized instructions can be found in Appendix D (Shinn, 2004). The examiners also followed the standardized guidelines for scoring CD, as formerly described. Appendix E includes a sample scoring probe comparable to those utilized by administrators to assist in counting CD.

Results

Table 3 details the descriptive statistics for the three M-CBM benchmarks and TCAP mathematics scaled scores by grade level. The data analytic process for this investigation involved two general steps. First, Pearson Product Moment Correlations

(Pearson's *r*) were calculated. These correlations examined the bivariate relationship between each of the three M-CBM assessments during the 2006-2007 school year and performance on the mathematics portion of the TCAP for each of the six participating grade levels.

Correlations between the TCAP and Fall, Winter, and Spring M-CBM benchmark assessments for all six participating grade levels are displayed within Table 4. The results indicate that correlations ranged from .24 to .49 (median $r = .43$), suggesting that the M-CBM benchmarking probes were moderately correlated with performance on the TCAP. Table 4 also details correlations determined among M-CBM probes for the Fall, Winter, and Spring for each of the participating grade levels. By grade, intercorrelations ranged from .47 to .73 (median $r = .57$). When M-CBMs were combined across grades and correlated, results indicated that the Fall, Winter, and Spring probes were all intercorrelated by at least $r = .60$.

The second part of the data analytic process utilized a linear multiple regression model with M-CBM scores from the Fall, Winter, and Spring benchmarks (all grade levels included) as the independent variables, and the scaled scores on the mathematics portion on the TCAP as the dependent variable. Together, the Fall, Winter, and Spring M-CBMs accounted for 26% of the overall variance in TCAP scores for the six participating grade levels. When grades 3-8 were analyzed separately, 11% to 28% (median = 25%) of the variance was explained. As indicated in Table 5, statistically significant predictors at each grade level differed due to the high multicollinearity among the Fall, Winter, and Spring M-CBM probes illustrated in Table 4. It should be noted

that one of the two grades with notably lower explained variance (eighth grade) had a large number of missing values for its Spring benchmarking period.

Discussion

This study evaluated the utility of M-CBM probes to predict performance on the TCAP, a statewide end-of-grade test. Specifically, this investigation explored the predictive validity of M-CBM benchmark probes at three time points throughout the year (Fall, Winter, and Spring) in reference to the TCAP for students enrolled in grades 3-8. Pearson Product Moment Correlations indicated M-CBM probes to be moderately correlated with performance on the mathematics portion of the TCAP across all six of the participating grade levels. Results from this analysis did not suggest a clear pattern of stronger correlational relationships based upon grade (e.g., older versus younger students) or for a particular benchmarking time (i.e., Fall, Winter, or Spring).

The pattern of results when utilizing a multiple linear regression analysis revealed statistically significant prediction for the TCAP for every grade level, explaining 11% to 28% of the variance. Two grades (fourth and eighth) did exhibit notably reduced prediction relative to the other four grades. Although an explanation for the lowered prediction rate for grade 4 is unclear, an explanation does present itself for grade 8. The lowered prediction rate for this grade level is likely due to the smaller sample size during the Spring benchmarking period. Overall, the results of this study extend previous work that supports the use of M-CBM as a predictor of performance on high-stakes tests (Helwig et al., 2002; Keller-Margulis et al., 2008).

The current study focused on two research questions. First, investigators wanted to determine to what degree the Fall, Winter, and Spring M-CBM benchmarking scores

correlated with scaled scores on the TCAP mathematics portion for each of the participating grade levels. Correlations ranged from .24 to .49 (median r = .43), suggesting M-CBM benchmarking probes are moderate predictors of performance on the mathematics portion of the TCAP. It was hypothesized that, at a minimum, the M-CBM scores would be moderately correlated with the TCAP. This hypothesis was supported by the data. For this investigation, it also was hypothesized that stronger relationships would be found for M-CBM and TCAP scores for the earlier grades participating in this study. This hypothesis was not supported by the data, since all correlations indicated moderate relationships between M-CBM and TCAP scores. Our results failed to converge with previous research (Foegen, 2008), which has found higher correlations between M-CBM measures and high-stakes test scores in the primary grades than in later grades. The results of the current study are similar to those obtained by other investigators who have reported, at a minimum, moderately high correlations between M-CBM scores and statewide high-stakes tests (Helwig et al., 2002). However, other investigations have reported stronger correlation coefficients between M-CBM and high-stakes test scores (Keller-Margulis et al., 2008) than those found within this investigation. The correlations between the M-CBM benchmarks and TCAP mathematics scaled scores may be lower than in other studies because of differences in the constructs that these two assessment tools measure. The TCAP measures a broad range of mathematical concepts, determined by state curriculum expectations for a particular grade. In contrast, M-CBM only measures math calculation skills. Therefore, M-CBM may only measure a particular subset of skills represented within the TCAP.

The second research question assessed during this investigation was whether or not there would be temporal differences across results. Specifically, the researchers wanted to determine if scores for one of the three M-CBM benchmarks (i.e., Fall, Winter, Spring) would prove to be a better predictor for TCAP performance than the other two benchmarks. It was hypothesized that Spring M-CBM scores would more effectively predict TCAP scores, since these were administered within a closer time frame during the school year. Although the regression analyses indicated that the Spring M-CBM was a statistically significant predictor for the TCAP for grades 5 and 7 only, results suggested that the Fall, Winter, and Spring M-CBMs were all moderately correlated with the TCAP. Given the high multicollinearity among the predictors, this does not contradict previous research finding stronger correlation coefficients between Spring M-CBM and a high-stakes test (Keller-Magulis et al., 2008). In the current study, all three of the M-CBMs were identified to be statistically significant predictors for various grade levels. In fact, the Fall M-CBM appears to at least predict as well as the Winter M-CBM and Spring M-CBM. This finding helps extend the evidence base supporting the use of M-CBM as a screening tool for identifying students at-risk for deficits in basic mathematical skills. Documenting the effectiveness of Fall M-CBMs as predictors of performance on high-stakes tests supports their utility in identifying students at-risk for academic failure earlier within the school year. This study clarifies that educators can indeed use tools, such as M-CBM, to screen students for academic deficits in the Fall. By doing so, we can better ensure that students will receive needed academic interventions earlier, and thereby address skills deficits in a timely manner.

Limitations and Directions for Future Research

Given the need for educators to provide early identification of students at academic risk and the value in using CBM measures to predict student performance on high-stakes tests, the results of this study are encouraging. However, as with any research, there are some limitations that should be noted. One major limitation is that the Spring M-CBM was given in May of 2007, and the TCAP was given in April of that same year. Since a linear regression model was used to show the ability of the three M-CBM measures to predict TCAP performance, it would have been more ideal for the Spring M-CBM to have been administered before the TCAP. However, this concern is somewhat attenuated, given the high levels of explained variance by any two M-CBM benchmarks (e.g., Fall and Winter).

Another possible limitation is that although the sample size of this current study was large in numbers, the characteristics of the sample were limited. The data for this study were gathered within a small, rural, school system that contained a low representation of students from diverse ethnic and SES backgrounds. Participants were mostly Caucasian, and the extent to which the current findings generalize to broader populations of students is unclear. Therefore, the norming population for the TCAP itself is more diverse and representative than the participants within this study.

Another limitation is that participants also represented a limited age group, since only grades 3-8 participated in the study. Grades 3 through 8 were chosen as participants because third grade is the initial year in which students begin to take the TCAP, and it is typical for high-stakes testing in other states across the nation to begin in grade 3. Future studies should strive to utilize longitudinal methods and extend the participant age range

to better assess the predictive validity of M-CBM in earlier grades (i.e., Kindergarten through second grade).

There are also limitations with regard to the criterion variable, the TCAP. High-stakes tests vary from state to state, making it difficult to determine how generalizable the results of this study may be for researchers investigating the prediction of student performance on high-stakes tests in other states. It also should be noted that questions included within the TCAP vary with each year. Furthermore, state officials appoint the cut scores used to determine student proficiency in an area on the TCAP, and these scores can change from one year to the next. Another limitation of the current study is attrition. Spring M-CBM scores were only available for a much smaller portion (i.e., around half) of students in grade 8. This smaller sample size was due to one middle school forgetting to administer Spring benchmark probes during the 2006-2007 school year.

Finally, another limitation is that the Fall, Winter, and Spring M-CBM measures were moderately intercorrelated with one another. This multicollinearity between the three M-CBM measures likely affected our ability to account more precisely for variance in TCAP mathematics scores for each of the participating grade levels.

Results of this study suggest directions for future research. More comprehensive multivariate analytic strategies (e.g., structural equation modeling) might yield a better understanding of the nature of the relation between M-CBM and statewide assessment measures, especially their relative predictive power across grade and the benchmarking time period. Early identification of students at-risk for academic failure is a central goal of RtI and the problem-solving model. Given this focus, it also would be beneficial for future researchers to utilize longitudinal methods to further assess the ability of M-CBM

benchmark scores to predict future student performance on high-stakes tests (i.e., gather student benchmarking data in Kindergarten and then compare results once these same students begin to take end-of-grade tests, typically beginning in third grade).

Implications for Practice

Despite these limitations, there are important implications of the results of this study. This investigation makes a distinctive contribution to the literature by investigating the relationship between M-CBM and performance on a statewide high-stakes test. Results do support the use of M-CBM within the current educational climate that promotes early identification of students at academic risk and response to intervention within a three-tiered model of services. It does seem that M-CBM measures skills relevant to those assessed by high-stakes tests and are valid tools to use for screening and monitoring students who are having difficulty with basic mathematic skills. This study adds to the literature identifying M-CBM as a reliable tool that educators can use early to identify students at-risk of academic failure. Since M-CBM is a quick and repeatable tool, educators also can utilize it to monitor student responsiveness to instruction targeting mathematical skill deficits. Results from this study also suggest that M-CBM does have the ability to predict student performance on a high-stakes test, which supports educators using M-CBM as a benchmarking tool to identify students at-risk for academic failure systematically within a school year. However, researchers should continue to investigate the possible reasons that prediction of student performance on high-stakes tests may vary based upon factors including grade level and the particular benchmarking period.

Conclusions

In summary, the present study adds to the research base of investigations that have found M-CBM probes to be a valid and reliable assessment of mathematical skills. Overall, the results suggest that M-CBM is strongly associated with performance in the TCAP for at least one of the three benchmarking periods. When utilized within an RtI model that focuses on identifying at-risk students and providing targeted academic intervention, M-CBM can offer systematic data to assist with instructional decision-making. M-CBM can help to guide educators toward the goal that all students demonstrate their mastery of grade level skills, as evidenced by passing end-of-year tests. With continued research, M-CBM may become an even stronger tool to assist educators in preparing students for high-stakes tests and therefore make instruction more focused and targeted toward mathematical skill deficits.

References

Arnold, V. A., & Smith, C. B. (1987). *Macmillan connections reading program*. New York: Macmillan Publishing Company.

Braden, J., & Tayrose, M. (2008). Best practices in educational accountability: High-stakes testing and educational reform. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 575-588). Bethesda, MD: National Association of School Psychologists.

Braden, J. P., & Schroeder, J. L. (2004). High-stakes testing and No Child Left Behind: Information and strategies for educators. *Helping children at home and school II: Handouts for families and educators*, pp.73-77. Retrieved from http://www.nasponline.org/communications/spawareness/highstakes.pdf

Brigance, A. (1999). *Comprehensive Inventory of Basic Skills* (rev. ed.). North Billerica, MA: Curriculum Associates, Inc.

Brown-Chidsey, R., Bronaugh, L., & McGraw, K. (2009). *RTI in the classroom: Guidelines and recipes for success*. New York, NY: The Guilford Press.

Carnoy, M., & Loeb, S. (2002). "Does external accountability affect student achievement? A cross-state analysis." *Educational Evaluation and Policy Analysis*, *4*, 287-301.

Chubb, J. E. (2005). *Within our reach: How America can educate every child*. Lanham, MD: Rowman & Littlefield.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading to predict student performance on statewide achievement tests. *Educational Assessment, 7*, 303-323.

Cusumano, D. L. (2007). Is it working? An overview of curriculum based measurement and its uses for assessing instructional, intervention, or program effectiveness. *The Behavior Analyst Today, 8*, 24-34.

Daly, E. J., & McCurdy M. (2002). Getting it right so they can get it right: An overview of the special series. *School Psychology Review, 31*, 453-458.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37,* 184-192.

Dworkin, A. G. (2005). The No Child Left Behind Act: Accountability, high-stakes testing, and roles for sociologists. *Sociology of Education, 78*, 170-174.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43,* 247-265.

Edformation. (2002). *AIMSweb progress monitoring and improvement system*. Available from http://www.aimsweb.com/

Elementary and Secondary Education Act of 1965, Pub. L. 89 -10, No. 79, Stat. 27, Ch. 70 (1965). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

Foegen, A. (2008). Progress monitoring in middle school mathematics. *Remedial and Special Education, 29,* 195-207.

Foegen, A., & Deno, S. L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education, 35*, 4-16.

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73*, 311-330.

Fuchs, L. S., Fuchs, D., & Hollenbeck, K. N. (2007). Extending responsiveness to intervention to mathematics at first and third grades. *Learning Disabilities Research & Practice, 22*, 13-24.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1998). *Monitoring Basic Skills Progress: Basic Math Computation* (2nd ed.) [Computer program]. Austin, TX: Pro-Ed.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring Basic Skills Progress: Basic Math Concepts and Applications* (2nd ed.) [Computer program]. Austin, TX: Pro-Ed.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38*, 293-304.

Good, R. H., III, & Kaminski, R. A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*, 326-336.

Grimm, K. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neural Psychology, 33*, 410-426.

Hasbrouck, J. E., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children, 24*, 41-44.

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education, 36*, 102-112.

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review, 31*, 514-528.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372-386.

Hoover, H., Dunbar, S., & Frisbie, D. (2005). *Iowa Tests of Basic Skills.* Chicago, IL: Riverside Publishing Company.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement.* New York, NY: Guilford Press.

Houghton Mifflin Basal Reading Series. (1989). *Journeys (grade 3). Discoveries (grade 2)*. Boston, MA: Author.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446 (2004). Retrieved from www.gpo.gov/fdsys/pkg/BILLS-108hr1350enr.pdf

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34*, 27-44.

Kaminski, R. A., & Good, R. H., III, (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*, 215-227.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*, 374-390.

Kelley, B. (2008). Best practices in using curriculum-based evaluation and math. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 419-437). Bethesda, MD: National Association of School Psychologists.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193-203.

Merrell, K. W., Ervin, R. A., & Gimpel, G. A. (2006). *School psychology for the 21st century: Foundations and practices.* New York, NY: The Guilford Press.

Minnesota Department of Education. (2003). *Minnesota Comprehensive Assessments: Grade 3 reading test specifications*. Roseville, MN: Author.

National Center for Education Statistics. (2009). The Nation's Report Card: Mathematics 2009 (NCES 2010–451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Mathematics Panel. (2007). National Mathematics Advisory Panel: Strengthening math education through research. Accessed November 18, 2009 from http://www.ed.gov/about/bdscomm/list/mathpanel/factsheet.html

The No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2001). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

Opfer, V. D., Henry, G. T., & Mashburn, A. J. (2008). The district effect: Systemic responses to high stakes accountability policies in six southern states. *American Journal of Education, 114,* 299-332.

Oregon Department of Education. (1999). *Assessment homepage*. www.ode.state.or.us//asmt/index.htm

Restori, A. F., Gresham, F. M., & Cook, C. R. (2008). Old habits die hard: Past and current issues pertaining to response-to-intervention. *The California School Psychologist, 13*, 67-78.

Reyna, V. F., & Brainerd, C. J. (2007). The importance of mathematics in health and human judgment: Numeracy, risk communication, and medical decision making. *Learning and Individual Differences, 17,* 147-159.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). Assessing response to instruction. In *Assessment in special and inclusive education.* (10th ed., pp. 629). Boston, MA & New York, NY: Houghton Mifflin Company.

Sattler, J. M. (2008). Challenges in assessing children: The context. In *Assessment of children: Cognitive foundations* (5th ed., pp. 22-54). La Mesa, CA: Jerome M. Sattler, Publisher, Inc.

Shinn, M. R. (2004). Administration and scoring of mathematics computation curriculum-based measurement (M-CBM) and math fact probes for use with AIMSweb. Bloomington, MN: NCS Pearson, Inc. Retrieved from http://www.aimsweb.com

Shinn, M. R. (2008). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 243-261). Bethesda, MD: National Association of School Psychologists.

Sibley, D., Biwer, D., & Hesch, A. (2001). *Establishing Curriculum-Based measurement oral reading fluency performance standards to predict success on local and state tests of reading.* Retrieved from ERIC database. (ED453527)

State of Tennessee Department of Education. (2007). *Tennessee Comprehensive Assessment Program Achievement Test: Parents' Guide to Understanding TCAP Achievement Test Results*. Monterey, CA: The McGraw-Hill Companies, Inc.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31,* 498-513.

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*, 363-382.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen A. (2007). CBM progress monitoring within a standards-based system: Introduction to the special series. *The Journal of Special Education, 41*, 66-67.

Table 1

Participant Numbers per Grade

Grade	Gender		Total
	Boys	Girls	
Third	117	107	224
Fourth	121	128	249
Fifth	162	162	324
Sixth	144	141	285
Seventh	168	157	325
Eighth	169	156	325
Total	881	851	1732

Table 2

Sample Characteristics of Participants

Ethnicity	Percentage
Caucasian	94.4
Hispanic	3.4
African American	1.4
Asian American	0.5
Native American	0.3

Table 3

Descriptive Statistics for M-CBM Benchmarks and TCAP by Grade Level

	Grade								
	3			4			5		
M-CBM	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Fall	204	14.68	6.22	229	30.79	11.60	293	28.70	11.59
Winter	209	23.30	7.71	231	38.39	13.75	308	36.72	15.72
Spring	209	27.27	9.26	237	47.32	17.42	292	41.39	16.48
TCAP	222	482.54	29.73	246	501.16	37.39	321	507.71	35.38
	6			7			8		
M-CBM	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Fall	268	25.19	9.73	306	32.76	12.59	308	35.59	13.97
Winter	263	32.62	13.16	301	40.86	15.48	307	46.52	16.56
Spring	264	29.91	13.31	227	39.11	16.15	154	42.24	17.66
TCAP	284	522.12	44.39	323	530.43	44.97	325	542.61	47.54

Note: M-CBM = Math Curriculum Based Measurement. *TCAP* = Tennessee Comprehensive Assessment Program. *M-CBM* scores are raw scores of digits correct per minute. The *TCAP* means and standard deviations are scaled scores.

Table 4

Correlations Among the Three M-CBM Benchmarks and TCAP by Grade Level

Grade	M-CBM	W	S	TCAP
3	F	.60	.53	.42
	W		.67	.48
	S			.41
4	F	.66	.52	.47
	W		.73	.32
	S			.24
5	F	.64	.67	.44
	W		.70	.43
	S			.46
6	F	.48	.48	.47
	W		.47	.49
	S			.39
7	F	.50	.48	.40
	W		.59	.42
	S			.45
8	F	.55	.57	.50
	W		.68	.34
	S			.26

Note: M-CBM = Math Curriculum Based Measurement. F = Fall; W = Winter; S = Spring. *TCAP* = Tennessee Comprehensive Assessment Program. *M-CBM* scores are raw scores of digits correct per minute. The *TCAP* means and standard deviations are scaled scores. *All correlations were statistically significant.

Table 5

Regression Analyses Predicting TCAP from M-CBM Benchmarks

	Beta				
Grade	Fall	Winter	Spring	F	R^2
3	.176*	.290**	.103	20.67***	.247
4	.377***	.070	.032	14.47***	.167
5	.137	.137	.283**	28.82***	.247
6	.257***	.282***	.125	33.87***	.284
7	.191**	.144	.277***	23.78***	.259
8	.213*	.072	.103	5.94***	.113

Note: *M-CBM* = Math Curriculum Based Measurement. *TCAP* = Tennessee Comprehensive Assessment Program. Regression parameters for M-CBM benchmarks are standardized betas. *p<.05, **p<.01, ***p<.001

Appendix A

5-13-2008

To Whom It May Concern:

The purpose of this letter is to grant researchers from Appalachian State University permission to disseminate data that were collected in our system during the 2003-2004, 2005-2006, 2006-2007, 2007-2008, and 2008-2009 school years. It is our understanding that these data were gathered as routine academic screenings and may include benchmark scores, progress monitoring scores, and other standardized test scores. Furthermore, we understand that no specific names of students, teachers, or schools will be communicated. Any other potential identifiers will be removed before these data are disseminated. We grant full permission to the researchers to disseminate these data via publication and presentation. Please do not hesitate to contact me if you have further questions.

Sincerely,

Mike Lowry
Director of Schools
Monroe County School System

Appendix B

Appalachian
STATE UNIVERSITY

INSTITUTIONAL REVIEW BOARD
Research and Graduate Studies
ASU Box 32068
Boone, NC 28608
828.262.2692
Web site: http://www.orsp.appstate.edu/compliance/irb/index.php
Email: irb@appstate.edu
Federalwide Assurance (FWA) #4801

To: Jamie Fearrington
Psychology
CAMPUS MAIL

From: _____
Jay W. Cranston, M.D., Chair, Institutional Review Board

Date: 6/02/2009

RE: Notice of IRB Exemption

Study #: 09-0260
Study Title: To What Degree Can High Stakes Test Scores Be Predicted By Math Curriculum-Based Measurement Performance?
Exemption Category: (4) Collection or Study of Existing Data, If Public or Unable to Identify Subjects

This submission has been reviewed by the above IRB Office and was determined to be exempt from further review according to the regulatory category cited above under 45 CFR 46.101(b). Should you change any aspect of the proposal, you must contact the IRB before implementing the changes to make sure the exempt status will continue. Otherwise, you will not need to apply for annual approval renewal. Please notify the IRB Office when you have completed the study.

CC:
Sara Reynolds, Psychology

Appendix C

AIMSweb® M-CBM Computation Benchmark #2 - Grade 3

Student Name: _____ **Grade:** _____ **Teacher Name:** _____

4 + 6	9 + 4	7 + 8	2 + 0	7 - 7	7 - 2
114 + 69	205 188 + 665	18 - 17	17 + 2	0 + 1	539 + 52
535 404 + 505	51 - 34	447 + 694	2 + 3	4 - 3	752 - 441
6 + 1	949 473 + 895	454 - 38	249 + 10	663 306 + 251	892 + 713
3 + 3	814 - 95	131 + 280	3 - 2	590 + 324	78 - 24
8 + 2	893 - 59	12 - 0	506 - 7	244 - 91	4 - 3

Appendix D

Math Curriculum-Based Measurement (M-CBM) Standard Directions

Grades 1-3 Probes

1. Students have an M-CBM probe and pencil.
2. Say to the student(s):

   "We're going to take a 2-minute math test. I want you to write your answers to several kinds of math problems. Some are addition and some are subtraction. Look at each problem carefully before you answer it.

   When I say 'BEGIN' write your answer to the FIRST problem (demonstrate by pointing) ***and work ACROSS the page. Then go to the next row.***

   Try to work EACH problem. If you come to one YOU REALLY DON'T KNOW HOW TO DO, put an 'X' through it and go to the next one.

   If you finish the first side, turn it over and continue working. Are there any questions? (Pause)***"***
3. Say ***"BEGIN"*** and start your stopwatch/timer.
4. If testing in groups, walk around and monitor students to ensure they are not skipping problems, are working across the page, and continue to write answers to the problems during the test time.

   If a student is excessively skipping problems they should know how to do, say to the student:
   "Try to work EACH problem. You can do this kind of problem so don't skip or put an 'X' over it."

   If a student is not working across the page, say to the student:
   "Work ACROSS the page. Try to work each problem in the row."

   If a student stops working before the test is done, say to the student:
   "Keep doing the best work you can."
5. At the end of 2 minutes, say ***"Stop. Put your pencils down."*** Monitor to ensure students stop working.

Appendix E

AIMSweb® M-CBM Computation Benchmark #2 - Grade 3 Answer Key

4 + 6 **10** (2)	9 + 4 **13** (2)	7 + 8 **15** (2)	2 + 0 **2** (1)	7 - 7 **0** (1)	7 - 2 **5** (1)	9 (9)
114 + 69 **183** (3)	205 188 + 665 **1058** (4)	18 - 17 **1** (1)	17 + 2 **19** (2)	0 + 1 **1** (1)	539 + 52 **591** (3)	14 (23)
535 404 + 505 **1444** (4)	51 - 34 **17** (2)	447 + 694 **1141** (4)	2 + 3 **5** (1)	4 - 3 **1** (1)	752 - 441 **311** (3)	15 (38)
6 + 1 **7** (1)	949 473 + 895 **2317** (4)	454 - 38 **416** (3)	249 + 10 **259** (3)	663 306 + 251 **1220** (4)	892 + 713 **1605** (4)	19 (57)
3 + 3 **6** (1)	814 - 95 **719** (3)	131 + 280 **411** (3)	3 - 2 **1** (1)	590 + 324 **914** (3)	78 - 24 **54** (2)	13 (70)
8 + 2 **10** (2)	893 - 59 **834** (3)	12 - 0 **12** (2)	506 - 7 **499** (3)	244 - 91 **153** (3)	4 - 3 **1** (1)	14 (84)

Vita

Sara Browning Reynolds was born in Virginia Beach, Virginia. She graduated from Hickory High School in Hickory, North Carolina in 1998. In May of 2002, she received her Bachelor of Science from Appalachian State University. After completing her undergraduate degree, she primarily worked as a Children's Case Manager for New River Behavioral Healthcare. In the fall of 2008, she began her graduate studies in School Psychology at Appalachian State University. In August 2010, she began her internship, working for Caldwell County Schools. During this time, she completed her internship requirements, while providing school psychology services to elementary, middle, and high school age students. Sara completed her Master of Arts degree in May of 2011 and continue her career as a School Psychologist. She currently resides in Boone, North Carolina with her husband and three dogs.